

**ORIGINAL
RESEARCH**

D.M.J. van den Heuvel
 V.H. ten Dam
 A.J.M. de Craen
 F. Admiraal-Behloul
 A.C.G.M. van Es
 W.M. Palm
 A. Spilt
 E.L.E.M. Bollen
 G.J. Blauw
 L. Launer
 R.G.J. Westendorp
 M.A. van Buchem
 on behalf of the
 PROSPER Study
 Group

Measuring Longitudinal White Matter Changes: Comparison of a Visual Rating Scale with a Volumetric Measurement

BACKGROUND AND PURPOSE: Detection of longitudinal changes in white matter hyperintensities (WMH) by using visual rating scales is problematic. We compared a widely used visual rating scale with a volumetric method to study longitudinal white matter changes.

METHODS: WMH were assessed with the visual Scheltens scale and a volumetric method in 100 elderly subjects aged 70–81 years for whom repetitive MR images were available with an interval of 33 (SD, 1.4) months. Reliability was determined by intraclass correlation coefficients. To examine the sensitivity of both the visual and volumetric method, we calculated Spearman rank correlations of WMH ratings and volume measurements with age.

RESULTS: Reliability of the visual rating scale was good, whereas reliability of the volumetric measurement was excellent. For baseline measurements of WMH, we found weaker associations between WMH and age when assessed with the visual scale ($r = 0.20$, $P = .045$) than with the volumetric method ($r = 0.31$, $P = .002$). Longitudinal evaluation of WMH assessments showed regression in 26% of the subjects when analyzed with the visual rating scale against 12% of the subjects when using volumetric measurements. Compared with the visual rating, the correlation between progression in WMH and age was twice as high when using the volumetric measurement ($r = 0.19$, $P = .062$ and $r = 0.39$, $P < .001$, respectively).

CONCLUSION: Volumetric measurements of WMH offer a more reliable, sensitive, and objective alternative to visual rating scales in studying longitudinal white matter changes.

White matter hyperintensities (WMH) are common findings on cerebral MR images of elderly people. Although the clinical significance has not yet been fully elucidated, WMH have been associated with cognitive impairment, gait dysfunction, and depressive symptoms.^{1–3} Several cerebrovascular risk factors and disorders have been associated with the presence of WMH.^{4,5} Moreover, older age is strongly related to WMH prevalence.^{4,5}

Longitudinal follow-up studies on the progression of WMH are essential for a thorough evaluation of the natural course, the clinical importance of the findings, and the effect of therapeutic interventions. However, several problems exist with evaluation of longitudinal MR imaging examinations. First, there are differences in image acquisition with time. Hence, images acquired at different time intervals might display data on WMH differently. Second, the actual measurement of change on repetitive white matter measurements might pose some problems. Until now, the presence of WMH has been primarily analyzed with visual rating scales.^{6–8} How-

ever, the detection of change on repetitive white matter measurements when using these scales is indeed problematic.⁹ Visual rating scales rely heavily on the human eye to detect changes in the cerebral white matter and describe WMH in a qualitative or semiquantitative way. Hence, intra- and inter-rater reliability is usually modest to low, and the rating scales are hampered by floor and ceiling effects. Automated or semi-automated volumetric lesion-detection methods largely overcome these shortcomings. Algorithm techniques replace the human eye, and 3D reconstruction and computation allow quantitative data on WMH. Thus, volumetric WMH measurements are more objective and reliable and provide exact measurements of WMH volume compared with the visual WMH ratings.¹⁰

So far, visual ratings and semiautomated volumetric measurements have not been compared directly as to their ability to assess progression in WMH—that is, the volumetric measurements have been used as a gold standard for the evaluation of the metric abilities of the visual rating scales.¹¹ However, the volumetric WMH measurements cannot be interpreted as such. External standards, like age or cognitive functioning, are needed to compare the sensitivity of both the visual and volumetric method.

We investigated the potential of both a widely used visual semiquantitative rating scale^{12,13} and an in house-developed semiautomated quantitative volumetric measurement to study longitudinal white matter changes. We compared both methods as to their reliability and sensitivity to detect change in white matter in advancing age.

Methods

Data were drawn from the nested MR imaging substudy of the Prospective Study of Pravastatin in the Elderly at Risk (PROSPER). Study

Received March 10, 2005; accepted after revision August 19.

From the Departments of Radiology (D.M.J.H., A.C.G.M.E., W.M.P., A.S., M.A.B.) and Gerontology and Geriatrics (V.H.D., A.J.M.C., G.J.B., R.G.J.W.) and the Division of Image Processing (F.A.-B.) and Neurology (E.L.E.M.B.), Leiden University Medical Center, the Netherlands; and the Division of Epidemiology, Demography Biostatistics (L.L.), the National Institute on Aging, the National Institutes of Health, Bethesda, Md.

This work was supported by an unrestricted research grant from Bristol-Myers Squibb.

The study was sponsored by an investigator-initiated grant from Bristol Myers-Squibb, Princeton, NJ. The sponsor had no role in the design, data collection, data analyses, data interpretation of the study, or writing of the paper. The authors declare the following arrangements with the sponsoring company and/or other companies making competing products: research support and travel grants, G.J. Blauw, M.A. van Buchem, E.L.E.M. Bollen, A.J.M. de Craen, and R.G.J. Westendorp.

Address correspondence to D.M.J. van den Heuvel, MSc, Department of Radiology C2-S, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, the Netherlands.

details have been described elsewhere.¹⁴ Of the 1100 eligible Dutch PROSPER participants, 554 consented and were included in the PROSPER MR imaging substudy. From these original 554 subjects, 100 randomly chosen subjects were evaluated in this study.

Image Acquisition. MR images were obtained from all subjects at baseline and after a mean follow-up of 33 (SD, 1.4) months. MR imaging was performed on a clinical MR imaging system operating at 1.5T field strength (Philips Medical Systems, Best, the Netherlands). With longitudinal MR imaging examinations, there is the problem of differences in section orientation with time—that is, both the angulations and positioning of the sections change between different MR imaging examinations. Hence, the presence of WMH on baseline and follow-up scans can differ because of changes in section orientation only. One can limit this effect by obtaining thin sections and by imaging sections without intersection gaps. Hence, we obtained dual fast spin-echo images with 48 contiguous 3-mm sections and no intersection gap (TR/TE, 3000/27/120 ms; matrix, 256 × 256; field of view, 220; acquisition percentage, 80%). To estimate the effect of section realignment with time, 8 participants were scanned and rescanned (with repositioning) in 1 MR imaging session. Short-term scan-rescan reliability for measurement of WMH was high (correlation coefficients [ICC] = 0.84).

Image Postprocessing. White matter hyperintensities were assessed with use of both the visual semiquantitative Scheltens scale and an in house–developed quantitative volumetric method. First, with the modified version of the semiquantitative Scheltens scale,^{12,13} WMH were traced on T2 proton density (PD)-weighted images in the periventricular and deep white matter areas (ratings for the basal ganglia and infratentorial areas were not used in this article) and rated according to size and number. The total score for the periventricular WMH ranged from 0 to 9 and for the deep WMH from 0 to 24. The total WMH score for the Scheltens scale for each participant thus ranged from 0 to 33, with a higher score indicating a larger amount of WMH. Second, quantification of WMH volume (milliliters) was performed by using in house–developed automated software (Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands).^{15,16} The exact method for the quantification of WMH has been described in detail elsewhere¹⁶ with the exception that in this article only PD- and T2-weighted images (dual MR images) were used for the automatic segmentation of WMH.

Briefly, our segmentation method combines information from dual MR images for volumetric analyses of WMH. The method is based on an established artificial intelligence technique: the Fuzzy inference system, which uses linguistic variables to classify a voxel. Each voxel is classified according to the voxel intensity (bright, medium-bright, dark) and according to the voxel position (white matter, intracranial). With the use of these linguistic values of voxel intensity and voxel position, fuzzy if-then rules were used to derive a label for every voxel. This label is then used to segment WMH from the rest of the brain with the use of another fuzzy inference rule: If voxel position is white matter and T2 intensity is “bright” and PD intensity is “bright,” then the segmented voxel is WMH. Hence, segmentations of WMH were generated automatically on dual MR images. The automatic segmentation still includes false-positives. Therefore, all images were edited manually to correct for incidental inclusion of CSF and gray matter. Moreover, fast fluid-attenuated inversion recovery hard copies were used as a reference to rule out other pathogenesis and entanglement of WMH with Virchow-Robin spaces. Infratentorial hyperintensities were excluded. The exact volumes of WMH were

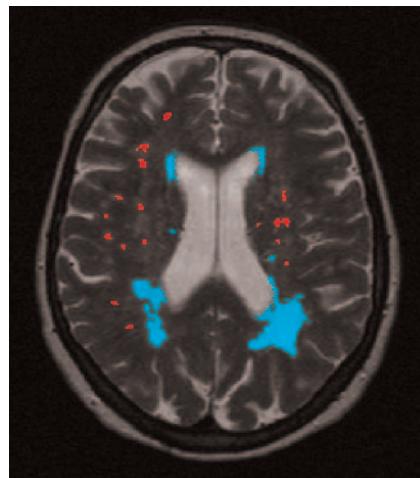


Fig 1. Computerized semiautomatic detection system for quantification of volumes of deep and periventricular WMH. Red indicates deep WMH and blue indicates periventricular WMH.

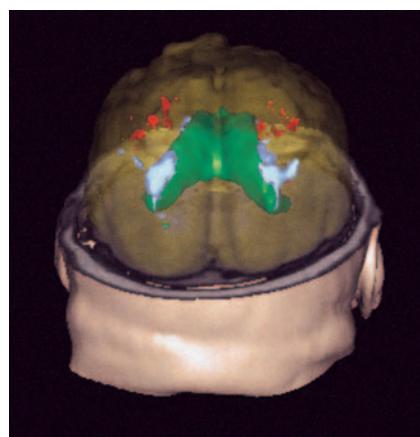


Fig 2. 3D representation of manually segmented deep WMH (red) and periventricular WMH (blue). The lateral ventricles are presented in green. The parenchyma is displayed in yellow.

calculated automatically. On average, 20 minutes per scan was spent on the manual editing of the segmented images. Examples of our semiautomatic WMH segmentation are given in Figs 1 and 2.

All raters were blinded to subject identity and had either much experience with use of the Scheltens scale (A.C.G.M.v.E., W.M.P.) or with volumetric assessment of WMH (V.H.t.D., D.M.J.v.d.H.). All ratings were performed within a time span of 1 month. Moreover, to prevent the possibility of over-reading WMH progression in a direct scan comparison setting, we analyzed baseline and follow-up MR images in random order. Fifteen MR images were segmented twice to assess the intra- and inter-rater reliability of the visual and volumetric WMH measurements.

Statistical Analyses. SPSS for Windows (release 11.0; SPSS, Chicago, Ill) was used for data analysis. The intra- and inter-rater reliability of the visual WMH ratings (ie, Scheltens score) and volumetric WMH measurements were determined by ICCs. ICCs equal to 0 reflected no agreement, whereas ICCs equal to 1 reflected total agreement. The associations between the visual rating and volumetric method were expressed as Spearman rank correlations. Moreover, to examine the sensitivity of both the visual and volumetric method, we calculated Spearman rank correlations of WMH ratings and volume measurements with age. The level of significance was set at $P < .05$.

Table 1: WMH measurement characteristics of the study sample

WMH measurement	Mean (SD)	Median (IQR)
Baseline		
Visual rating scale* (points)	10.5 (7.8)	9.0 (5.0–14.8)
Volumetric measurement (mL)	6.5 (11.8)	1.4 (0.4–6.4)
Follow-up		
Visual rating scale* (points)	11.5 (7.9)	9.0 (6.0–16.0)
Volumetric measurement (mL)	8.8 (14.1)	2.8 (0.6–10.2)
Increase		
Visual rating scale* (points)	1.0 (3.5)	1.0 (−1.0–3.0)
Volumetric measurement (mL)	2.3 (3.6)	1.1 (0.1–2.7)

Note:—* Scheltens scale. WMH indicates white matter hyperintensities; IQR, interquartile range.

Results

The mean age of the 100 randomly selected participants in this study was 74.5 years ($SD = 2.9$). Forty-one percent were women. Visual ratings and volumetric measurements of WMH of the study participants are presented in Table 1.

Intra- and inter-rater reliability of the visual WMH rating was good ($ICC = 0.83$ and 0.74 , respectively), whereas the intra- and inter-rater reliability of the volumetric WMH measurement was excellent ($ICC = 1.00$ and 0.99 , respectively).

To measure the strength of the linear relationship between the visual rating and the volumetric measurement, we estimated the correlation between both WMH measurements. There was a high correlation between the visual ratings and volumetric measurements of WMH at baseline ($r = 0.89$, $P = .000$) and at end of follow-up ($r = 0.91$, $P = .000$). However, the correlation of the estimates of longitudinal WMH progression between the visual ratings and volumetric measurements was substantially lower ($r = 0.29$, $P = .003$).

To examine the sensitivity of visual ratings and volumetric measurements of WMH, we evaluated the associations of either type of WMH measurement at baseline with age. As expected, both the visual ratings and the volumetric measurements of WMH were significantly associated with age, though the association with age was weaker for the visual ratings ($r = 0.20$, $P = .045$) than for the volumetric measurements ($r = 0.31$, $P = .002$).

A longitudinal evaluation of the visual rating scale showed that 26% of participants had a regression in WMH, whereas 12% regressed when the volumetric method was used (Fig 3, Table 2). Furthermore, when the progression in WMH with age was evaluated, we found that the correlation of the volumetric WMH measurement was twice as high as that of the visual rating scale ($r = 0.19$, $P = .062$ and $r = 0.39$, $P < .001$, respectively).

Discussion

It has been postulated that volumetric measurements of WMH might be the optimal solution for the evaluation of WMH progression.¹⁰ We are the first to actually compare the potential of a semiquantitative visual rating scale and a semi-automated quantitative volumetric method to study longitudinal changes in WMH. Our data demonstrate that the volumetric measurement was more reliable and more sensitive for the evaluation of white matter changes with time.

Two studies have assessed the sensitivity and reliability of some widely used visual rating scales for measuring white matter changes.^{9,11} Both studies concluded that although the ex-

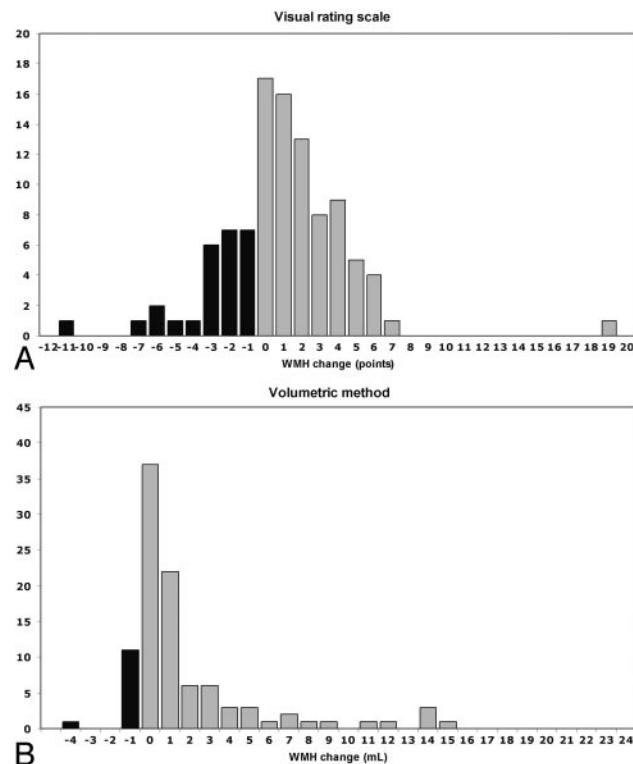


Fig 3. Graph shows longitudinal evaluation of WMH change in 100 elderly subjects by using (A) a visual rating scale (ie, Scheltens scale) and (B) a volumetric method. Black indicates the number of subjects showing WMH regression at the end of follow-up; gray indicates the number of subjects showing no change in WMH or WMH progression at the end of follow-up.

Table 2: Frequency of WMH regression in study sample (n = 100) after 3 years of follow-up

WMH Regression	Frequency
Visual*	
1 point	7
2–4 points	14
5–10 points	4
>10 points	1
Total	26
Volumetric	
≤0.1 mL	4
>0.1–0.2 mL	3
>0.2–0.3 mL	2
>0.3–0.5 mL	2
>0.5 mL	1
Total	12

Note:—* Scheltens scale. WMH indicates white matter hyperintensities.

isting visual rating scales suffice for WMH measurements in cross-sectional studies, they are not sufficiently reliable and sensitive for measuring white matter changes in longitudinal analyses. Our present findings strongly support this view.

The inability of visual rating scales to measure change in WMH on longitudinal MR images is best illustrated by looking at the unexpected finding of regression of WMH in 26% of our study population when we used the visual rating scale. With aging, WMH are found to increase and not to decrease with time.^{4,5} Hence, the reported regression of WMH with time is probably an artifact due to measurement error. This artifact also occurred when using the volumetric method. However, only 12% of our

study population showed regression of WMH when we used the volumetric method. Moreover, the relative size of the measurement error was far smaller for the volumetric method compared with the visual rating scale.

Recently, Prins et al¹¹ presented a scale that was specifically designed for measuring WMH changes with time. However, their scale remains a visual rating scale; therefore, it inherently encompasses most of the difficulties of the existing visual rating scales. First, all visual rating scales have user interference and subjectivity and are, therefore, less reliable when compared with a volumetric method. Second, visual rating scales have a reduced sensitivity because they are prone to floor and ceiling effects due to their relatively large measurement units (ie, categories) and fixed upper limits. Therefore, progression in WMH beyond the defined upper limit cannot be detected. Furthermore, subtle changes in WMH are neglected if the progression is within category limits. Moreover, in visual rating scales with only a few categories (ie, qualitative scales), the extent of the progression in WMH is completely disregarded. Hence, true progression will be under- or overestimated. In contrast, volumetric measurements have small measurement units (ie, voxels) and no upper limit. Volumetric methods thus allow an unrestricted estimate of WMH volumes and change in WMH volume with time.

The main strength of this study is the large series of baseline and follow-up scans that were analyzed by using both a visual rating scale and a volumetric method. Moreover, so far, visual ratings and volumetric measurements have not been compared directly in their ability to assess progression in WMH with the use of an external standard. On the other hand, this feature is also one of the limitations of our study. In fact, there is no accepted gold standard for the assessment of WMH. However, because age has unequivocally been related to the presence of WMH in the literature,^{4,5} we argued that it could well serve as an external standard. The sensitivity of both the visual rating scale and volumetric measurement would presumably translate well into a closer association with age.

Conclusion

Our data suggest that volumetric measurements of WMH offer a more reliable, sensitive, and objective alternative to visual rating scales in studying longitudinal white matter changes. In addition, volumetric measurements of WMH enable comparison between various studies on the same topic. Although extensive time requirements for volumetric WMH quantification are often reasons given to limit their use, highly sophisticated and time-efficient methods of semiautomated volumetric WMH are now within reach.

Acknowledgments

We thank H. Olofsen, MD, for technical assistance and N. Schmitz, MD, and W. M. van der Flier, MD, for critical reading and valuable comments.

PROSPER Study Group

Executive Committee: (Glasgow) J. Shepherd (chairman and principal investigator), S.M. Cobbe, I. Ford, A. Gaw, P.W. Macfarlane, C.J. Packard, D.J. Stott; (Leiden) G.J. Blauw (principal investigator), E.L.E.M. Bollen, A.M. Kamper, R.G.J. Westendorp; (Cork) M.B. Murphy (principal investigator), B.M. Buckley, M. Hyland, I.J. Perry.

End Point Committee: S.M. Cobbe (chairman), J.W. Jukema, P.W. Macfarlane, A.E. Meinders, D.J. Stott, B.J. Sweeney, C. Twomey.

Data and Safety Monitoring Committee: W.V. Brown (chairman), H.C. Diener, J. Feely, I. Ford (nonvoting), T. Pearson, S. Pocock, P.A. van Zwieten.

References

1. De Groot JC, De Leeuw FE, Oudkerk M, et al. Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. *Ann Neurol* 2000;47:145–51
2. Whitman GT, Tang Y, Lin A, et al. A prospective study of cerebral white matter abnormalities in older people with gait dysfunction. *Neurology* 2001;57:990–94
3. Barber R, Scheltens P, Ghokar A, et al. White matter lesions on magnetic resonance imaging in dementia with Lewy bodies, Alzheimer's disease, vascular dementia, and normal aging. *J Neurol Neurosurg Psychiatry* 1999;67:66–72
4. Awad IA, Spetzler RF, Hodak JA, et al. Incidental subcortical lesions identified on magnetic resonance imaging in the elderly. I. Correlation with age and cerebrovascular risk factors. *Stroke* 1986;17:1084–89
5. Ylikoski A, Erkinjuntti T, Raininko R, et al. White matter hyperintensities on MRI in the neurologically nondiseased elderly: analysis of cohorts of consecutive subjects aged 55 to 85 years living at home. *Stroke* 1995;26:1171–77
6. Schmidt R, Fazekas F, Kapeller P, et al. MRI white matter hyperintensities: three-year follow-up of the Austrian Stroke Prevention Study. *Neurology* 1999;53:132–39
7. Veldink JH, Scheltens P, Jonker C, et al. Progression of cerebral white matter hyperintensities on MRI is related to diastolic blood pressure. *Neurology* 1998;51:319–20
8. Wahlund LO, Almkvist O, Basun H, et al. MRI in successful aging: a 5-year follow-up study from the eighth to ninth decade of life. *Magn Reson Imaging* 1996;14:601–08
9. Kapeller P, Barber R, Vermeulen RJ, et al. Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement, and correlations with quantitative measurements. *Stroke* 2003;34:441–45
10. Payne ME, Fetzer DL, MacFall JR, et al. Development of a semi-automated method for quantification of MRI gray and white matter lesions in geriatric subjects. *Psychiatry Res* 2002;115:63–77
11. Prins ND, van Straaten EC, van Dijk EJ, et al. Measuring progression of cerebral white matter lesions on MRI: visual rating and volumetrics. *Neurology* 2004;62:1533–39
12. Scheltens P, Barkhof F, Leys D, et al. A semiquantitative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *J Neurol Sci* 1993;114:7–12
13. Garde E, Mortensen EL, Krabbe K, et al. Relation between age-related decline in intelligence and cerebral white-matter hyperintensities in healthy octogenarians: a longitudinal study. *Lancet* 2000;356:628–34
14. Shepherd J, Blauw GJ, Murphy MB, et al. The design of a prospective study of Pravastatin in the Elderly at Risk (PROSPER): PROSPER Study Group—PROSpective Study of Pravastatin in the Elderly at Risk. *Am J Cardiol* 1999;84:1192–97
15. van der Flier WM, Middelkoop HA, Weverling-Rijnsburger AW, et al. Interaction of medial temporal lobe atrophy and white matter hyperintensities in AD. *Neurology* 2004;62:1862–64
16. Admiraal-Behloul F, van den Heuvel DMJ, Olofsen H, et al. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. *Neuroimage* 2005;28:607–17. Epub 2005 Aug 29